LA-UR-18-24540
Approved for public release; distribution is unlimited.

Title:	Hierarchical Linear Regression
Author(s):	Wallstrom, Timothy Clarke Higdon, David Mitchell
Intended for:	Report
Issued:	2019-01-04 (rev.1)

Hierarchical Linear Regression

Timothy C. Wallstrom
T-4, Los Alamos National Laboratory
tcw@lanl.gov

David M. Higdon
CCS-6, Los Alamos National Laboratory
Biocomplexity Institute, Virginia Tech
dhigdon@vbi.vt.edu

LA-UR-18-24540

December 21, 2018

Abstract

We describe an algorithm for making inferences in a hierarchical linear model. The model allows different slopes and intercepts for different groups, and ties together the different slopes by representing them as samples from a higher level distribution. (The intercepts are treated as statistically independent.) Our model and code permit many options for the variance of this hierarchical distribution for the slopes, including gamma distributions on the precision, t-distributions on the standard error, pooled slopes (all equal), and independent slopes. The algorithm is described in the paper; an implementation of the algorithm in R is available on request.

1 Introduction

We consider inference in a two-level linear regression model, in which the slopes are drawn from a normal hyperprior:

$$y_{ij} \sim N(\alpha_i + \beta_i \cdot t_{ij}, \sigma^2) \qquad (i = 1, \ldots, g, j = 1, \ldots, n_i). \tag{1}$$

$$\beta_i \sim N(\beta_0, \sigma_\beta^2). \tag{2}$$

Here, α_i and β_i are the intercept and slope for the ith group, y_{ij} is the jth measurement in group i, taken at time t_{ij}, and σ^2 is the variance of the measurements. We assume g groups, with n_i measurements in group i. In the hierarchical model, the parameters β_0 and σ_β are assigned prior probability distributions, and the posterior of all the parameters is inferred through Bayesian inference.

The model is useful when the distribution of slopes is expected to be exchangeable though not necessarily independent [3]. The degree of similarity between the slopes is determined by the distribution of σ_β. If the prior forces σ_β to be zero, then the slopes are the same in all groups, and the model is sometimes called the "pooled model." If the prior forces σ_β to be very large, then the slopes are statistically independent. Either of these cases can, of course, be implemented directly in simpler non-hierarchical models.

Unless there is a great deal of data, the inferences will be quite sensitive to the choice of prior on σ_β. There is generally little prior information about σ_β so we often want to use a noninformative prior. There are a number of different proposals for noninformative priors; see [3, Sec. 5.7.3] for a discussion and further references. Also, we may want to use informative priors sometimes. It is often useful to test the sensitivity of one's inferences the choice of prior, or to parameters in the prior, and for this reason, it is useful to be able to easily test a variety of different priors.

The purpose of this paper is to describe a hierarchical regression algorithm that is capable of using a variety of different priors for σ_β, including (i) the root-inverse-gamma prior (corresponding to the gamma prior on the precision, $\lambda_\beta \equiv 1/\sigma_\beta^2$); (ii) the "flat" prior, recommended as a noninformative prior in [1]; the limiting cases of a (iii) pooled model and an (iv) independent model; and (v) the half-t priors, also recommended as noninformative choices in [1]. The code is available on request.

2 Definition of priors

We give formulas for some recommended priors.

2.1 Root inverse gamma prior

The density of the root-inverse-gamma (RIG) prior [3, Sec. 2.6.6] is

$$p_{\text{RIG}}(\sigma \mid a, b) = \frac{2b^a}{\Gamma(a)} \sigma^{-2a-1} e^{-b/\sigma^2}. \tag{3}$$

The root-inverse-gamma distribution on σ is equivalent to the gamma distribution on the precision, $\lambda = 1/\sigma^2$, where

$$p_{\text{Gamma}}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}.$$

a is called the *shape parameter* in both distributions. b is called the *inverse scale parameter* in the gamma distribution. In the RIG distribution, σ scales as $\sqrt{b}$.

2.2 The flat prior

The flat prior can be realized as an RIG$(-\frac{1}{2}, 0)$ prior. The prior is improper, but the likelihood is proper when there are at least three groups [1, p. 21].

2.3 Pooled and independent priors

When discussing limiting cases, it is somewhat easier to work in terms of the precision λ than the standard deviation, because the mean and standard deviation of the gamma distribution are available in closed form, as a/b and $\sqrt{a}/b$, respectively. The pooled prior corresponds to the limit $\lambda \to \infty$. It can be approximated by $\lambda_\beta \sim \text{Gamma}(a, b)$ with a large and $b = 1$, say. The independent prior corresponds to the limit $\lambda \to 0$. This prior can be implemented by taking large b and $a = 1$, say.

2.4 Half-t prior

The density of the "half-t" distribution, $|t_\nu|$, is

$$p_{|t_\nu|}(\theta|s^2) = \frac{2\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{1}{2}\nu)\sqrt{\nu\pi}s} \left(1 + \frac{1}{\nu}\left(\frac{\theta}{s}\right)^2\right)^{-\frac{1}{2}(\nu+1)} \qquad (\theta > 0). \tag{4}$$

We are particularly interested in the case $\nu = 1$, the "half-Cauchy" distribution, with density

$$p_{\text{HC}}(\theta|s^2) = \frac{2}{\pi} \frac{s}{s^2 + \theta^2} \qquad (\theta > 0). \tag{5}$$

3 Model

In order to have a single code implementing all these priors, it is helpful to write $\beta_i = \beta_0 + \delta_i$, where $\delta_i = \xi\eta_i$. The model is

$$y_{ij} \sim N(\alpha_i + (\beta_0 + \xi\eta_i) \cdot t_{ij}, \sigma^2) \qquad (i = 1, \ldots, gj = 1, \ldots, n_i). \tag{6}$$

$$\eta_i \sim N(0, \sigma_\eta^2). \tag{7}$$

A graphical representation of the model is shown in Fig 3.

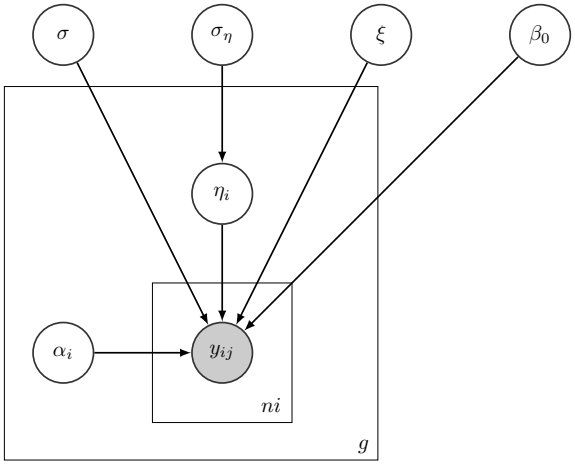

Figure 1: Graphical representation of hierarchical linear model with parameter expansion. Variables are identified with nodes. The distribution of a node is conditional on the nodes pointing to it. Rectangles are "plates," which are duplicated the number of times indicated in the lower right corner.

Writing $\beta_i = \beta_0 + \delta_i$ is simply a means of separating off the mean. Rather than writing $\beta_i \sim N(\beta_0, \sigma_\beta^2)$, we can equivalently write $\beta_i = \beta_0 + \delta_i$, where $\delta_i \sim N(0, \sigma_\delta^2)$, provided we assign σ_δ the same distribution as σ_β.

Writing $\delta = \xi\eta$ is technique known as "parameter expansion," [1], which permits the efficient evaluation of new priors for σ_δ. If $\xi = 1$ then $\eta_i = \delta_i$, and nothing is changed. If, on the other hand, we take

$$\xi \sim N(0, 1) \tag{8}$$

$$\sigma_\eta \sim \text{RIG}(\tfrac{1}{2}\nu, \tfrac{1}{2}\nu s^2), \tag{9}$$

then it can be shown that

$$\sigma_\delta \sim |t_\nu|(s^2),$$

as can be confirmed by marginalizing over ξ (see Appendix). The advantage of this expansion is that the priors for both ξ and σ_η are now conditionally conjugate, i.e., they are of the same form as the distributions of the same parameters, conditioned on all the other parameters and the data (the so-called "full-conditionals"), so they can both be updated using Gibbs sampling (see below). Thus, the new model provides a way of simulating the half-Cauchy prior on σ_δ, or more generally the half-t, using Gibbs sampling. Parameter expansion also modifies the geometry of parameter space to provide better mixing for the Markov chain.

4 Inference

We describe inference in the parameter-expanded model. Inference is by Gibbs sampling, which is a particularly efficient special case of the Metropolis algorithm in which the sampling distribution is available analytically and the acceptance probability is one. The full set of variables are α, η, β_0, ξ, $\lambda \equiv 1/\sigma_\eta^2$, σ, and y. The joint distribution is of the form

$$p(\alpha, \eta, \beta_0, \xi, \lambda, \sigma, y) = \\ p(y \mid \alpha, \eta, \xi, \beta_0, \sigma, \lambda)\, p(\eta \mid \lambda)\, p(\alpha)\, p(\sigma)\, p(\lambda)\, p(\xi)\, p(\beta_0). \quad (10)$$

We update (α, η, β_0) as one block, following ideas of [2].

4.1 α, η, β_0

We have

$$p(\alpha, \eta, \beta_0 \mid \xi, \sigma, \lambda, y) \propto p(\alpha, \eta, \beta_0, y \mid \xi, \sigma, \lambda) \quad (11)$$
$$= p(y \mid \alpha, \eta, \beta_0, \xi, \sigma, \lambda)\, p(\alpha, \eta, \beta_0 \mid \xi, \sigma, \lambda). \quad (12)$$

We consider the two terms on the rhs of Eq. (12) separately.

For the first term, suppose that the y_{ij} are ordered into a vector y_k, where $k = k(i, j)$. According to the model Eq. (4.4),

$$p(y \mid \alpha, \eta, \beta_0, \xi, \sigma, \lambda) = p(y \mid \phi, \sigma),$$

where $\phi = (\alpha, \xi\eta, \beta_0)$. Let X be the design matrix for Y, i.e.,

$$Y \sim N(X\phi, \Sigma), \quad (13)$$

where

$$\Sigma = \mathrm{diag}(\sigma^2).$$

The model for y, Eq. (4.4), is

$$p(y \mid \phi, \sigma) = \left(\det \frac{W(\sigma)}{2\pi}\right)^{1/2} \exp\{-\tfrac{1}{2}(y - X\phi)^T W(\sigma)(y - X\phi)\}, \quad (14)$$

where $W(\sigma) = \Sigma^{-1}$.

For the second term,

$$p(\alpha, \eta, \beta_0 \mid \xi, \sigma, \lambda) = \frac{p(\alpha, \eta, \beta_0, \xi, \sigma, \lambda)}{p(\xi, \sigma, \lambda)} \\ = p(\eta \mid \lambda),$$

as is easily derived from the joint distribution Eq. (10).

From Eq. (12), we have

$$p(\phi \mid \xi, \sigma, \lambda, y) \propto \exp\left\{-\tfrac{1}{2}(y - X\phi)^T W(\sigma)(y - X\phi) - \tfrac{1}{2}\lambda\xi^{-2}\phi^T R\phi\right\}, \quad (15)$$

where $R \equiv \text{diag}(\vec{0}_g, \vec{1}_g, 0)$. The exponent is now quadratic in ϕ. Completing the square, we get

$$\boxed{\phi \mid y, \sigma, \xi, \lambda \sim N(\hat{\phi}, \Sigma_\phi)}, \tag{16}$$

where

$$\Sigma_\phi(\sigma, \lambda, \xi) \equiv \left(X^T W(\sigma) X + \lambda \xi^{-2} R\right)^{-1}$$
$$\hat{\phi}(\sigma, \lambda, \xi, y) \equiv \Sigma_\phi X^T W(\sigma) y.$$

We use a t-superscript to indicate discrete time in the Markov Chain. To update (α, η, β_0), we sample

$$\phi^* \sim \phi \mid \lambda^{t-1}, \sigma^{t-1}, \xi^{t-1}, y,$$

and writing $\phi^* = (\alpha^*, \delta^*, \beta_0^*)$, set

$$\alpha^t \leftarrow \alpha^*$$
$$\eta^t \leftarrow \delta^* / \xi^{t-1}$$
$$\beta_0^t \leftarrow \beta_0^*.$$

4.2 ξ

We have

$$\begin{aligned} \xi \mid \cdots &\propto e^{-\xi^2/2} \prod_{ij} p(y_{ij} \mid \alpha_j, \eta_j, \beta_0, \xi), \\ &\propto e^{-f(\xi)}, \end{aligned}$$

where

$$f(\xi) = -\tfrac{1}{2} \sum_{ij} w_{ij} \left(y_{ij} - (\alpha_j + (\beta_0 + \xi \eta_j) t_{ij})\right)^2 - \tfrac{1}{2} \xi^2.$$

Here and below, $\cdots$ represents all the other variables (parameters and data). It is straightforward to infer that

$$\boxed{\xi \mid \cdots \sim N(\hat{\xi}, \sigma_\xi^2),} \tag{17}$$

where

$$\hat{\xi}(\alpha, \eta, \beta_0) = \frac{\sum_{ij} w_{ij} \eta_j t_{ij} \left(y_{ij} - \alpha_j - \beta_0 t_{ij}\right)}{1 + \sum_{ij} w_{ij} (\eta_j t_{ij})^2}$$
$$\sigma_\xi^2(\eta) = \left(1 + \sum_{ij} w_{ij} (\eta_j t_{ij})^2\right)^{-1}.$$

We generate ξ^t via Gibbs' sampling:

$$\xi^t \sim N\left(\hat{\xi}(\alpha^t, \eta^t, \beta_0^t), \sigma_\xi^2(\eta^t)\right)$$

4.3 λ

It can be seen from the graph that λ is only dependent on the η_j. Thus,

$$\begin{aligned} p(\lambda \mid \cdots) &= p(\lambda \mid \vec{\eta}) \\ &\propto p(\vec{\eta} \mid \lambda)\, p(\lambda). \end{aligned}$$

We have

$$p(\vec{\eta} \mid \lambda)\, p(\lambda) \propto \lambda^{g/2} e^{-\frac{1}{2}\lambda\, \vec{\eta}^2} \lambda^{a-1} e^{-b\lambda},$$

so

$$\boxed{\lambda|\vec{\eta} \sim \text{Gamma}\left(a + \tfrac{1}{2}g, b + \tfrac{1}{2}\vec{\eta}^2\right)}.$$

We generate λ^t by setting

$$\lambda^t \sim \lambda \mid \eta^t.$$

4.4 σ

With $p(\sigma) = 1/\sigma$, we have

$$p(\sigma \mid \dots) \propto \frac{1}{\sigma^{n+1}} \exp\left[-\frac{1}{2}\frac{(y - X\phi)^T (y - X\phi)}{\sigma^2}\right],$$

so that

$$\boxed{\sigma \mid y, \phi \sim \text{RIG}\left(\tfrac{1}{2}n, \tfrac{1}{2}\,(y - X\phi)^T (y - X\phi)\right).}$$

We generate σ^t by taking

$$\sigma^t \sim \sigma \mid y, \phi^t,$$

where $\phi^t = (\alpha^t, \xi^t \eta^t, \beta_0^t)$.

The variance σ^2 is sometimes broken into two parts, representing measurement uncertainty $\sigma^2_{m,ij}$, depending on the data point and assumed known, and a variation σ_r^2, which does not depend on the datapoint and is inferred statistically. The algorithm described here is easily generalized to a model in which the residual variation is modeled separately. The data equation is then

$$y_{ij} \sim N(\alpha_i + \beta_i \cdot t_{ij}, \sigma^2_{m,ij} + \sigma_r^2) \qquad (i = 1, \dots, g; i = 1, \dots, n_i).,$$

We assume a prior $p(\sigma_r)$ on σ_r. We then replace Eq. (13) with

$$Y \sim N(X\phi, \Sigma(\sigma_r)), \tag{18}$$

where

$$\Sigma(\sigma_r)_{kk} = \text{diag}(\sigma_r^2 + \sigma^2_{m,i(k),j(k)})),$$

and $W(\sigma_r) \equiv \Sigma^{-1}(\sigma_r)$. (Recall that $k(i,j)$ is an ordering of the datapoints y_{ij}. $i(k)$ and $j(k)$ are the inverse functions back to the group and instance labels.)

With these changes, all of the above equations go through with σ replaced by σ_r, but σ_r needs to be updated using a Metropolis step. Let "..." denote the time t value of all parameters other than σ_r, and compute

$$\begin{aligned} r^t &= \frac{p(\sigma_r^*, \ldots \mid y)}{p(\sigma_r, \ldots \mid y)} \\ &= \frac{p(\sigma_r^*, \ldots, y)}{p(\sigma_r, \ldots, y)} \\ &= \frac{p(y \mid \phi^t, \sigma_r^*)}{p(y \mid \phi^t, \sigma_r)} \frac{p(\sigma_r^*)}{p(\sigma_r)}. \end{aligned}$$

We set

$$\sigma_r^t = \begin{cases} \sigma_r^* & r^t > u \\ \sigma_r^{t-1} & r^t \le u, \end{cases} \tag{19}$$

where u is a uniform random variable on $[0, 1]$. In practice, we compare $\log r^t$ and $\log u$; for reference,

$$\begin{aligned} \log r^t = -\tfrac{1}{2}\mathrm{res}(\phi^t, \sigma_r^*) - \tfrac{1}{2}\log|\Sigma(\sigma_r^*)| + \log p(\sigma_r^*) \\ + \tfrac{1}{2}\mathrm{res}(\phi^t, \sigma_r^t) + \tfrac{1}{2}\log|\Sigma(\sigma_r^t)| - \log p(\sigma_r^t), \end{aligned}$$

where $\mathrm{res}(\phi^t, \sigma_r)$, the residual sum of squares is

$$\mathrm{res}(\phi^t, \sigma_r) = \sum_{ij} \frac{\left(y_{ij} - (X\phi^t)_{k(i,j)}\right)^2}{\sigma_r^2 + \sigma_{ij}^2}.$$

Author contributions

The original algorithm and code for the Gamma distribution was written by DMH. The extension to the t-distribution, and the writing of this paper, is due to TCW.

Acknowledgments

This work was supported by the Advanced Certification Campaign at Los Alamos National Laboratory, and by the Department of Energy under contract DE-AC52-06NA25396.

References

[1] Andrew Gelman, *Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)*, Bayesian Analysis **1** (2006), no. 3, 515–534.

[2] Daniel J Sargent, James S Hodges, and Bradley P Carlin, *Structured Markov Chain Monte Carlo*, Journal of Computational and Graphical Statistics **9** (2000), no. 2, 217–234.

[3] David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley & Sons, 2004.

Appendix: Parameter Expansion

For simplicity, we consider a hierarchical model with constant group-level effects, rather than the linear regression model; the calculations should be essentially the same. Consider two models:

$$\begin{aligned} y_{ij} &\sim N(\alpha_i, \sigma_{ij}^2) \\ \alpha_i &\sim N(0, \sigma^2) \\ \sigma &\sim |t_\nu|(s^2), \end{aligned}$$

and the parameter-expanded model,

$$\begin{aligned} y_{ij} &\sim N(\xi\eta_i, \sigma_{ij}^2) \\ \xi &\sim N(0, 1) \\ \eta_i &\sim N(0, \sigma^2) \\ \sigma^2 &\sim \text{IG}(\tfrac{1}{2}\nu, \tfrac{1}{2}\nu s^2). \end{aligned}$$

We show that these models are identical, in the sense that the joint distribution of $\{y_{ij}\}$ is the same. We do so by writing out the distribution for the parameter-expanded model and marginalizing over ξ, which yields the distribution of the original model.

$$\begin{aligned} p(y) &\propto \iiint \prod_{ij} e^{-\frac{(y_{ij}-\xi\eta_i)^2}{2\sigma_{ij}^2}} \cdot \prod_i \frac{e^{-\frac{1}{2}\eta_i^2/\sigma^2}}{\sigma} \cdot e^{-\frac{1}{2}\xi^2} \, \frac{1}{(\sigma^2)^{\frac{1}{2}\nu+1}} e^{-\frac{\nu s^2}{2\sigma^2}} \, (\textstyle\prod_i d\eta_i) \, d\xi \, d(\sigma^2) \\ &\propto \iiint \prod_{ij} e^{-\frac{(y_{ij}-\xi\eta_i)^2}{2\sigma_{ij}^2}} \cdot \prod_i \frac{e^{-\frac{1}{2}\alpha_i^2/\tau^2}}{\tau} \cdot e^{-\frac{1}{2}\xi^2} \, \frac{\xi^{\nu+2}}{(\tau^2)^{\frac{1}{2}\nu+1}} e^{-\frac{\xi^2}{2}(\frac{\nu s^2}{\tau^2})} \, \frac{(\prod_i d\alpha_i) \, d\xi \, d(\sigma^2)}{\xi^2} \end{aligned}$$

Separating off the terms in ξ, we get

$$\int \xi^\nu \exp\left[-\frac{\xi^2}{2}\left(1 + \frac{\nu s^2}{\tau^2}\right)\right] d\xi = \int \xi^\nu e^{-\frac{1}{2}(A\xi)^2} \, d\xi \propto \frac{1}{A^{\nu+1}},$$

where

$$A = \left(1 + \frac{\nu s^2}{\tau^2}\right)^{1/2}.$$

Thus,

$$p(y) \propto \iint \prod_{ij} e^{-\frac{(y_{ij}-\xi\eta_i)^2}{2\sigma_{ij}^2}} \cdot \prod_i \frac{e^{-\frac{1}{2}\alpha_i^2/\tau^2}}{\tau} \cdot \frac{1}{(\tau^2)^{(\nu+1)/2}} \frac{1}{A^{\nu+1}} \prod_i d\alpha_i \cdot d\tau$$

$$p(y) \propto \iint \prod_{ij} e^{-\frac{(y_{ij}-\xi\eta_i)^2}{2\sigma_{ij}^2}} \cdot \prod_i \frac{e^{-\frac{1}{2}\alpha_i^2/\tau^2}}{\tau} \cdot \frac{1}{(\tau^2+\nu s^2)^{(\nu+1)/2}} \prod_i d\alpha_i \cdot d\tau,$$

which proves the assertion.

Gelman [1] has a simpler argument: he observes that $\sigma_\alpha = |\xi|\sigma_\eta$, which is the ratio of a half normal and the square-root of a Gamma-distributed variable, and thus a t. However, the individual α_i are correlated, and we do not see how to extend his argument to show that the correlation between the $\xi\eta_i$ is the same as the correlation between the α_i.